Journal of Risk (39–54) Volume 9/Number 2, Winter 2006/07

Quality control of risk measures: backtesting VAR models

Victor H. de la Pena*
Department of Statistics, Columbia University, Room 1027, Mail Code 4690, 1255 Amsterdam Avenue, New York, NY 10027, USA

Ricardo Rivera
State of New York Banking Department and New York University, One State Street, New York, NY 10004, USA

Jesus Ruiz-Mata
Lehmann Brothers, 745 7th Avenue, Second Floor, New York, NY, 10019, USA

This paper introduces a new statistical approach to assessing the quality of risk measures: quality control of risk measures (QCRM). The approach is applied to the problem of backtesting value-at-risk (VAR) models. VAR models are used to predict the maximum likely losses in a bank's portfolio at a specified confidence level and time horizon. The widely accepted VAR backtesting procedure outlined by the Basel Committee for Banking Supervision controls the probability of rejecting the model when the model is correct. A drawback of the Basel approach is its limited power to control the probability of accepting an incorrect VAR model. By exploiting the binomial structure of the testing problem, QCRM provides a more balanced testing procedure, which results in a uniform reduction of the probability of accepting a wrong model.

QCRM consists of three elements: the first is a new hypothesis-testing problem in which the null and alternative hypotheses are exchanged to control the probability of accepting an inaccurate model. The second element is a new approach for comparing the power of the QCRM and Basel tests in terms of the probability of rejecting correct and incorrect models. The third element involves the use of the technique of pivoting the cumulative distribution function to obtain one-sided confidence intervals for the probability of an exception.

The use of these confidence intervals results in new acceptance/rejection regions for tests of the VAR model. We compare these to ones commonly used in the financial literature.

1 Introduction

Banking regulators and risk managers of financial institutions often wish to assess the adequacy of a model. They frequently test the null hypothesis that the model is correct against an alternative hypothesis that the model is incorrect. In this paper we present an approach to assist regulators and risk managers in this

*Corresponding author.
The first author acknowledges support from NSF grant DMS 5-24517.

task. We illustrate the approach by presenting an application to the problem of backtesting value-at-risk (VAR) models.

The $(1-\delta)100\%$ VAR number is the $(1-\delta) \times 100$ percentile of the distribution of the portfolio losses for a specified time horizon. VAR backtesting is the process by which financial institutions periodically compare daily profits and losses with the model-generated risk measures to gauge the accuracy of their VAR models.

In 1996 the Basel Committee for Banking Supervision ("Basel") developed a framework for backtesting the internal models used to calculate regulatory capital for market risk (see Basel (1996)). Alternative statistical evaluation methodologies have been proposed. For example, the evaluation based on the binomial distribution is discussed by Kupiec (1995); the interval forecast method proposed by Christoffersen (1998); the distribution forecast method by Crnkovic and Drachman (1996); and the magnitude loss function method by Lopez (1996).

We start with the following question:

- *What are regulators and bank managers really doing when they apply the Basel VAR backtesting methodology?*

The Basel backtesting procedure tests the null hypothesis that the bank's VAR model predicts losses accurately against the alternative hypothesis that the model predicts losses incorrectly:

$$H_0^B: \text{VAR model is correct vs. } H_1^B: \text{VAR model is incorrect} \qquad (1)$$

The test statistic used is based on the number of exceptions generated by the VAR model. For a given trading day, an exception occurs when the loss exceeds the model-based VAR.

The test postulates that the probability of an exception, p, is equal to 0.01 and tests it against the alternative hypothesis that the probability of an exception is greater than 0.01. The test is based on the number of exceptions in 250 trading days. The test rejects the VAR model if the number of exceptions is greater than or equal to 10 and accepts the model otherwise.

When evaluating statistical tests, it is common practice to examine their type I and II error rates. Under the Basel backtesting procedure, the type I error rate is the probability of rejecting the VAR model when the model is correct, while the type II error rate is the probability of accepting the model when the model is incorrect.

The Basel backtesting procedure is designed for controlling the α level, the probability of rejecting the VAR model when the model is correct. For this test, the α level is the probability that number of exceptions, out of 250 daily observations, is greater than or equal to 10, when the probability of an exception is $p = 0.01$. The α level of the test is 0.0003 (or 0.03%).

Although the test establishes a very conservative threshold for controlling the type I error rate, it is not designed to control the type II error rate. Therefore,

it does not control the probability of accepting the VAR model when the model is incorrect.

This drawback is pointed out on page 5 of Basel (1996): "The Committee of course recognizes that tests of this type are limited in their power to distinguish an accurate model from an inaccurate model".

To address this issue we introduce an alternative approach to the Basel backtesting procedure. We call this methodology quality control of risk measures (QCRM). Its goal is to enhance the ability of the test to reject an incorrect model. It consists of three elements: first, the test introduces a new hypothesis testing problem in which the null and alternative hypotheses are exchanged. The goal is to control the probability of accepting a wrong model. The second element consists of a new definition of the power of the tests that allows the comparison of QCRM and Basel backtesting procedures. The third element involves the use of a technique to obtain accurate estimates of the acceptance/rejection regions.

The new hypothesis testing problem is:

$$H_0^Q\text{: VAR model is incorrect vs. } H_1^Q\text{: VAR model is correct} \qquad (2)$$

Under QCRM, the acceptance of the null hypothesis is equivalent to the rejection of the VAR model. The model is accepted when the null hypothesis in (2) is rejected and, hence, when there is overwhelming evidence supporting the model.

A type I error for QCRM occurs when the VAR model is accepted but the model is incorrect, and a type II error happens when the model is rejected although the model is correct. Table 1 shows correct and incorrect decisions when the null and the alternative hypotheses are true.

The following relationship is important for understanding the properties of the QCRM test:

Probability of rejecting the VAR model when the model is incorrect

= 1 – Probability of accepting the VAR model when it is incorrect.

Under QCRM the probability of accepting the VAR model when the model is incorrect is set at level α: this means that the procedure controls the type 1 error rate. By setting the α level to a small value ($\leq 1\%$), QCRM guarantees a high probability ($\geq 99\%$) of rejecting a wrong model. Once this level is fixed, the use

TABLE 1 Type I and II errors for QCRM.

True hypothesis	Decision 1: VAR model is rejected	Decision 2: VAR model is accepted
Model correct	***Type II error***	Correct assessment
Model incorrect	Correct assessment	***Type I error***

of the Neyman–Pearson lemma provides a uniformly most powerful (UMP) test that minimizes the type II error (or the probability of rejecting a correct model).

We introduce an approach for comparing the power of the QCRM and Basel tests in terms of the probability of rejecting correct and incorrect models. Using this approach, it is shown that, relative to the Basel procedure, QCRM provides a balanced trade-off between type I and type II error rates (see Table 3). Table 5 and Figure 1 provide numerical comparisons of the power of the tests.

Similar to the Basel backtesting procedure, QCRM establishes the acceptance/rejection zones based on probabilistic arguments (see (8)). More precisely, QCRM uses the technique of pivoting the true cumulative distribution function of the observations to find one-sided confidence intervals, which provides the new three zones for accepting/rejecting the VAR model.

The paper is organized as follows: Section 2 reviews the Basel methodology for testing VAR models. Section 3 introduces the basic idea of QCRM and the new hypothesis testing problem. Section 3.2 presents the statistical techniques to compute appropriate coverage confidence intervals for the true (unknown) probability of an exception, p. The proposed rules for accepting/rejecting the VAR model are obtained in Section 3.3. The approach for comparing the relative power of the two tests is then addressed in Section 3.4. Finally, Section 4 summarizes our work. Some technical proofs are presented in Appendices A and B at the end of the paper.

2 Basel VAR backtesting methodology

The VAR backtesting methodology compares actual losses against model-based VAR calculations. Let the losses observed in every trading day be denoted by

$$L_1, L_2, \ldots, L_n \tag{3}$$

Let p be the probability of an exception on a given trading day. The computed $(1-p) \times 100\%$ one-day VAR numbers for the corresponding trading days are

$$V_0^1, V_1^2, \ldots, V_{n-1}^n, \cdots \tag{4}$$

where V_{i-1}^i denotes the VAR calculated for day i using the information available on day $i-1$. Define the indicator random variables

$$Y_i = 1_{\{L \geq V_{i-1}\}}, \quad i \geq 1 \tag{5}$$

such that $Y_i = 1$ if $L_i \geq V_{i-1}^i$ and $Y_i = 0$ otherwise. Hence, an exception occurs on trading day i if and only if $Y_i = 1$, ie, there is a loss exceeding VAR.

The Basel backtesting procedure assumes that the conditional probability of an exception at day i, given the information up to day $i-1$, is a fixed value, ie, $P(Y_i | I_{i-1}) = p$. This implies that the indicator variables $Y_1, Y_2, \ldots$, have fixed

conditional probabilities. In Appendix A we show that the assumption of constant conditional probability is a necessary and sufficient condition for independence of the indicator variables $Y_i, \ldots, Y_n$.

The Basel backtesting procedure implicitly tests the following hypothesis:

$$H_0^B : p = p_0 \quad \text{vs.} \quad H_A : p > p_0 \tag{6}$$

based on the sample of $n = 250$ observations $Y_1, \ldots, Y_{250}$ (where $p_0 = 0.01$). The Basel procedure uses a p_0 value of 1%. This hypothesis is tested against the alternative hypothesis that the probability of an exception is greater than p_0.

The Neyman–Pearson lemma provides a uniformly most powerful (UMP) test with rejection region of the form

$$R = \{\#\text{ of exceptions} \geq k\}$$

where k is a threshold obtained from the desired probability of the type I error, ie,

$$P\left(\#\text{ of exceptions} \geq k \,|\, p = p_0\right) = P\left(R \,|\, H_0\right) = \alpha \tag{7}$$

When $k = 10$, Basel backtesting gives $\alpha = 0.003$ or 0.3%, which implies that with a 99.997% probability the VAR model will not be rejected when the model is correct (ie, $p = 0.01$).

Based on the number of exceptions of the VAR model, Basel defines three zones: The green zone consists of four or fewer exceptions, and in this case the VAR model is assessed as correct. The yellow zone includes five to nine exceptions, and the accuracy of the VAR model is questioned. The red zone corresponds to 10 or more exceptions, and the model is rejected.

The limits for each of the zones can be linked to a probabilistic statement through the cumulative distribution of the number of the exceptions. The green zone extends from 0 to 95% of the cumulative probability distribution (cdf) of the observed exceptions. The yellow zone starts where the cdf of the observed exceptions exceeds 95% up to 99.97%. The red zone corresponds to values of probability greater than the 99.99% percentile of the cdf of the observed exceptions. For example, using the variables in (3), we compute:

$$\begin{aligned} &P\left(\sum_{i=1}^{250} Y_i \leq 4 \,|\, p = 0.01\right) = 0.8922, \quad P\left(\sum_{i=1}^{250} Y_i \leq 5 \,|\, p = 0.01\right) = 0.9588, \\ &P\left(\sum_{i=1}^{250} Y_i \leq 9 \,|\, p = 0.01\right) = 0.9997, \quad P\left(\sum_{i=1}^{250} Y_i \leq 10 \,|\, p = 0.01\right) = 0.9999 \end{aligned} \tag{8}$$

Note that these probabilities are calculated assuming that the true parameter p is equal to 0.01.

3 Quality control of risk measures

3.1 New hypothesis testing problem

QCRM starts with the hypothesis that the VAR model is incorrect and then tests this against the alternative hypothesis that the VAR model is correct. Accepting the null hypothesis then implies the rejection of the VAR model, while rejecting the null hypothesis leads to the acceptance of the model.

Let us assume that the probability of an exception is p_0 when the VAR model is correct and is some unknown value p_1 when the model is incorrect. More specifically, the rival hypotheses are

$$\mathrm{H}_0^Q: p > p_1 \quad \text{vs.} \quad \mathrm{H}_1^Q: p \leq p_0 \tag{9}$$

where p is the unknown probability of an exception. Note that, using monotonic properties of tests, the hypothesis test in (9) is equivalent to saying $\mathrm{H}_0^Q : p = p_1$ vs. $\mathrm{H}_1^Q : p \leq p_0$. This results from the fact that the likelihood ratio used in developing the test is a monotone function of the parameter p (see Equation (8) in Appendix B).

The test in (9) above can be seen as a quality control problem in which regulators and risk managers assess the quality of the risk model and reject it if the proportion of exceptions (model failures) is statistically larger than an acceptable threshold; eg, $p_0 = 0.01$.

To construct the test, we assume that the observations are independent so that the number of exceptions

$$S_n = Y_1 + \ldots + Y_n = \sum_{i=1}^{250} Y_i = X \tag{10}$$

follows a binomial distribution with parameters n and p. The statistic S_n is also a sufficient statistic for p, and this means that it contains all the information available about the parameter p.

In Appendix B we show that the statistic S_n has a cumulative distribution function that decreases in the parameter p. Therefore, using the theorem developed by Karlin-Rubin (see Casella and Berger (2002)), the test that rejects $\mathrm{H}_0^Q : p > p_1$ vs. $\mathrm{H}_1^Q : p \leq p_0$ if and only if $\{S_n \leq s(p_1)\}$ is uniformly most powerful level α test, where $\alpha = P_{p_1}(S_n \leq s(p_1))$. Intuitively, the unknown probability p cannot be large if the number of exceptions is small.

Under the new hypotheses testing problem, QCRM controls the probability of accepting the VAR model when the model is incorrect. By setting α to a small level, QCRM safeguards against incorrectly accepting a false model.

3.2 Optimal confidence intervals for hypothesis testing

In this section we provide the theoretical underpinnings of the QCRM acceptance/rejection regions. More specifically, we use the correspondence between

tests of hypotheses and interval estimation to obtain optimal confidence intervals.

For each sample value $Y = (Y_1, \ldots, Y_n)$, define the confidence region as follows:

$$C(Y) = \{p_1 : Y \in A(p_1)\} \tag{11}$$

where $A(p_1)$ is the acceptance region for $p = p_1$ against $p = p_0$.

The traditional method for developing a one-sided confidence interval for p uses the normal approximation for the sample proportion, $\hat{p} = x/n$ and computes the lower boundary as

$$p_L(x, \alpha) = \hat{p} - z_\alpha \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where z_α is the $(1 - \alpha) \times 100\%$ quantile of the standard normal distribution. The value is used to obtain a right one-sided $(1 - \alpha) \times 100\%$ confidence interval $(p_L(x, \alpha), 1]$ for p_1.

However, as indicated in Brown, Cai and DasGupta (2001), the intervals have deficiencies in coverage probability originating from the discreteness of the distribution, even when n is large, when np is small (as when $n = 250$ and $p = 0.01$).

To overcome this problem, we use a technique in which we pivot the true cdf of the observations to compute one-sided intervals with a confidence level at least as large as the desired confidence level (see Casella and Berger (2002)). One advantage of this technique is that it gives confidence intervals with the desired confidence level.[1]

Consider the tail probability of the sample proportion of exceptions, $\hat{p}$, in a period of n days. The interval $(p_L(x, \alpha), 1]$ is a right one-sided interval of p with a coverage level greater than or equal to $(1 - \alpha) \times 100\%$, where $p_L(x, \alpha)$ is the smallest value of p which satisfies the following:

$$1 - F_x(x) = P\left(\hat{P} \geq \frac{x}{n} \,\middle|\, p_L(x, \alpha)\right) \leq \alpha \tag{12}$$

As shown below, Equation (12) provides the basis for the development of QCRM's acceptance/rejection regions.

3.3 Probabilistic-based rules for accepting/rejecting VAR models

The proposed rules for accepting/rejecting VAR models are computed by inverting the rejection region of the test in (8) with level of significance α (see Equation (12)).

When inverting the rejection region, we obtain a right one-sided confidence interval, $(p_L(X, \alpha), 1]$, for p with a coverage level greater than or equal to $(1 - \alpha) \times 100\%$, ie, $P(p \in (p_L(X, \alpha), 1]) = 1 - \alpha$. Therefore, the test certifies that the model is correct, with at least $(1 - \alpha) \times 100\%$ confidence, every time p_0

[1] Technical proofs of pivoting the true cdf are provided in Appendix B.

TABLE 2 95% and 99% right-sided confidence intervals for the probability of an exception after observing k exceptions in $n = 250$ trading days.

Regions	95%	99%
Green		
$k = 1$	[0, 1]	[0, 1]
$k = 2$	(0.000202, 1]	(0.000039, 1]
$k = 3$	(0.0033, 1]	(0.0017, 1]
$k = 4$	(0.0055, 1]	(0.0033, 1]
$k = 5$	(0.0079, 1]	(0.0051, 1]
Yellow		
$k = 6$	(0.0105, 1]	(0.0072, 1]
$k = 7$	(0.0132, 1]	(0.0094, 1]
Red		
$k = 8$	(0.0160, 1]	(0.0117, 1]
$k = 9$	(0.0189, 1]	(0.0142, 1]
$k = 10$	(0.219, 1]	(0.0167, 1]

belongs to the interval $(p_L(X, \alpha), 1]$, or alternatively, when p_0 does not belong to $(0, p_L(X, \alpha)]$.

Note the following property of these one-sided confidence intervals:

$$(p_L(X, \alpha), 1] \subset (p_L(X, \alpha^*), 1] \quad \text{if } \alpha^* < \alpha \tag{13}$$

By analogy to the Basel supervisory framework, QCRM defines the following new zones:

- *New green zone* The VAR model is certified as correct if p_0 is in the 95% one-side confidence interval for p, $(p_L(X, 0.05), 1]$.
- *New yellow zone* When p_0 is not in the one-sided 95% confidence interval but is in the 99% one-sided confidence interval for p, $(p_L(X, 0.01), 1]$, then the validity of the model is questioned.
- *New red zone* If p_0 is not in the 99% confidence interval for p, $(p_L(X, 0.01), 1]$, then the VAR model is rejected.

We employ a Newton–Raphson algorithm to obtain a solution to Equation (12). Table 2 presents the 95% and 99% right one-sided confidence intervals for the probability of an exception after observing k exceptions ($0 \leq k \leq 10$).

Using Table 2 and our criteria for defining the zones, the 99% VAR model is certified when zero to five exceptions are observed. The model is questioned when six or seven exceptions are observed. Finally, the model is rejected when eight or more exceptions are realized.

Therefore, the proposed three zones using QCRM are:

- New green zone = {0 to 5 exceptions}
- New yellow zone = {6 or 7 exceptions}
- New red zone = {8 or more exceptions}

3.4 A new approach for comparing powers: a tale of two powers

As noted, QCRM controls the probability of accepting the model when the model is incorrect at a low level α for the test. Therefore, it rejects the null hypothesis that the model is inaccurate when there is overwhelming evidence supporting the model (ie, when $p \leq 0.01$). This results in the statistical validation of the model.

To compare the results of the QCRM and Basel procedures, we look at the relative power of the tests. Since the tests are different in nature (the null and alternative hypotheses are different), it is not possible to directly compare their powers. However, by defining the power function of the tests in terms of the probability of rejecting correct and incorrect models, we are able to make appropriate comparisons.

The proposed approach for comparing test powers is based on the probability of rejecting correct and incorrect models. The first step is to evaluate type I and type II error rates. Using the results of the tests for 250 observations and a 1% VAR, the type I and type II error rates are as shown in Table 3.

The next step is to define the power function for QCRM, $\beta^{Q}(p)$, as:

- For $p > 0.01$, $\beta^{Q}(p) = P(\text{Rejecting the model} \mid \text{Model is incorrect})$
  $= 1 - P(\text{Accepting the model} \mid \text{Model is incorrect})$
  $= 1 - P(X \leq 7 \mid \text{Given } p) = P(X \geq 8 \mid \text{Given } p)$, and

- For $p \leq 0.01$, $\beta^{Q}(p) = P(\text{Rejecting the model} \mid \text{Model is correct})$
  $= P(X \geq 8 \mid \text{Given } p)$.

Likewise, the power function for Basel, $\beta^{B}(p)$, is:

- For $p > 0.01$, $\beta^{B}(p) = P(\text{Rejecting the model} \mid \text{Model is incorrect})$
  $= 1 - P(\text{Accepting the model} \mid \text{Model is incorrect})$
  $= 1 - P(X \leq 9 \mid \text{Given } p) = P(X \geq 10 \mid \text{Given } p)$, and

- For $p \leq 0.01$, $\beta^{B}(p) = P(\text{Rejecting the model} \mid \text{Model is correct})$
  $= P(X \geq 10 \mid \text{Given } p)$.

Table 5 compares the power of the tests under the new approach. Relative to Basel, QCRM significantly increases the probability of rejecting a wrong model. The relative power gain is partially offset by an increase of the probability of rejecting a correct model.

TABLE 3 Type I and II error rates for the QCRM test.

	Decision with QCRM test	
True hypothesis	**Accept H_0^Q**	**Reject H_0^Q**
H_0^Q: $p > 0.01$	OK	Type I error = $P(X \leq 7 \mid p > 0.01)$
H_A^Q: $p \leq 0.01$	Type II error = $P(X \geq 8 \mid p \leq 0.01) = [0, 0.004]$	OK

TABLE 4 Type I and II error rates for Basel test.

	Basel test decision	
True hypothesis	**Accept H_0^B**	**Reject H_0^B**
H_0^B: $p = 0.01$	OK	Type I error = $P(X \geq 10 \mid p = 0.01) = 0.0003$
H_A^B: $p > 0.01$	Type II error = $P(X \leq 9 \mid p > 0.01)$	OK

TABLE 5 Powers of Basel and QCRM tests.

	Probability of rejecting the model when it is:	
Test	**Correct**	**Incorrect**
Basel	Less than 0.0003*	$P(X \geq 10 \mid \text{Given } p > 0.01)$
QCRM	Less than 0.004	$P(X \geq 8 \mid \text{Given } p > 0.01)$†

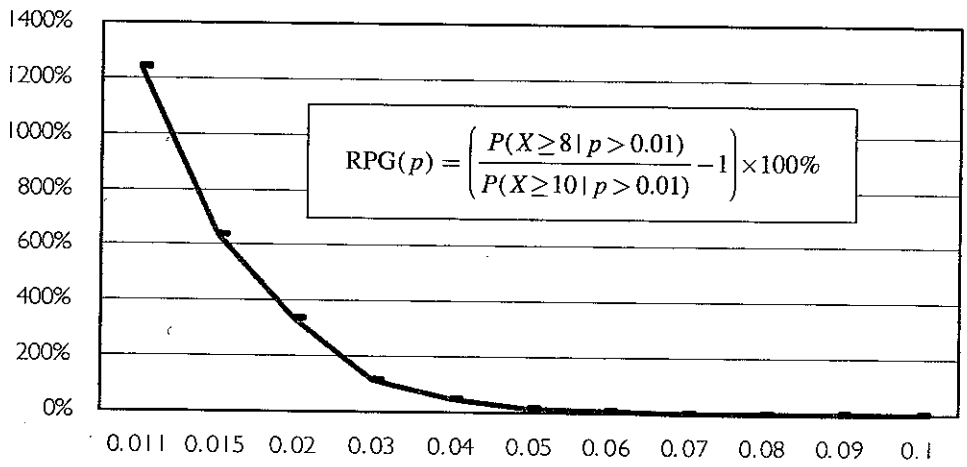
*Assumes a composite null hypothesis for the Basel test with $p \leq 0.01$.
†QCRM rejects the VAR model when the number of exceptions is equal to or greater than eight for alternative values of p greater 0.01. See Section 3.3 for a proof of this.

The Basel test is very conservative in controlling the probability of rejecting a correct model (ie, 0.03%); however, it does so at the cost of increasing, relative to QCRM, the probability of accepting a wrong model.

On the other hand, the QCRM test provides a balanced trade-off between type I and type II errors since it uses a less extreme probability of rejecting a correct model. As a result, QCRM delivers significant gains in power. For example, Figure 1 shows the percentage gains of QCRM over Basel, calculated as percentages of the respective probabilities of rejecting a wrong model, for values of the (unknown) probabilities of an exception greater than 0.01.

Note that QCRM outperforms the Basel test, especially for parameter values closer to 0.01. For instance, there are rate gains in the range of 637% to 115% for p values between 0.015 and 0.03, respectively.

FIGURE 1 Relative power gain (RPG) curve.



4 Conclusions

In this paper we have developed an approach for validating risk models and applied it to the problem of validating a value-at-risk model. The main contribution of our approach is the development of a probabilistic-based test to control the probability of accepting an incorrect VAR model. As a result, a model is validated when there is enough evidence to support it.

A second contribution is the introduction of an approach for comparing the powers of the tests in terms of the probability of rejecting correct and incorrect models. Using this approach, we have shown that the QCRM procedure delivers significant gains in power, relative to the Basel procedure, in terms of the probability of rejecting a wrong model.

The third element is the calculation of improved estimates of the probability of acceptance/rejection regions. The traditional method for obtaining a one-sided confidence interval for the parameter p uses the normal approximation for the sample proportion. However, the intervals have deficiencies in coverage probability originating from the discreteness of the distribution. We have obtained confidence intervals with the exact desired confidence level by pivoting the cumulative distribution function of the number of exceptions.

Our results show that a VAR model should be certified as correct when no more than five exceptions are observed and should be rejected if eight or more exceptions are observed in 250 trading days.

The proposed methodology can be extended to address other model validation problems within or outside the financial world.

Appendix A: Independence

In this section we prove the statement that Bernoulli random variables with fixed conditional probabilities must be independent.

Consider a sequence of Bernoulli random variables $X_1,\ldots,X_n$, $P(X_i=1)=p$ and $P(X_i=0)=1-p$, and assume that for every n there exists a fixed p such that

$$P\left(X_{n+1}=x_{n+1}\,|\,X_0=x_0,\ldots,x_n\right)=p^{x_{n+1}}(1-p)^{1-x_{n+1}} \tag{A1}$$

We now prove by induction that (A1) implies that $X_1,X_2,\ldots$ are independent Bernoulli random variables with parameter p. Consider the case $n=0$; then (A1) implies that

$$\begin{aligned}P\left(X_0=x_0,X_1=x_1\right)&=P\left(X_1=x_1\,|\,X_0=x_0\right)P\left(X_0=x_0\right)\\&=p^{x_1}(1-p)^{1-x_1}P\left(X_0=x_0\right)\\&=P\left(X_0=x_0\right)P\left(X_1=x_1\right)\end{aligned} \tag{A2}$$

which implies that X_0 and X_1 are independent Bernoulli random variables. Let us assume now that the first n random variables $X_0,\ldots,X_n$ are independent Bernoulli random variables with parameter p; then, using the induction hypothesis and (A1) gives

$$\begin{aligned}&P\left(X_0=x_0,\ldots,X_n=x_n,X_{n+1}=x_{n+1}\right)\\&\quad=P\left(X_{n+1}=x_{n+1}\,|\,X_0=x_0,\ldots,X_n=x_n\right)P\left(X_0=x_0,\ldots,X_n=x_n\right)\\&\quad=p^{x_{n+1}}(1-p)^{1-x_{n+1}}p^{x_n}(1-p)^{1-x_n}\ldots p^{x_0}(1-p)^{1-x_0}\\&\quad=p^{\sum_{i=0}^{n+1}x_i}(1-P)^{n-\sum_{i=0}^{n+1}x_i}\\&\quad=\prod_{i=0}^{n+1}P(X_i=x_i)\end{aligned}$$

For the general case, consider a set of indexes $i_1<i_2<\ldots<i_m$; then, using the law of total probabilities,

$$\begin{aligned}P\left(X_{i_1}=x_{i_1},\ldots,X_{i_m}=x_{i_m}\right)&=\sum_{x_i's=0,\,x_i\neq x_{ji},j=1,\ldots,m}^{1}P\left(X_1=x_1,\ldots,X_{i_m}=x_{i_m}\right)\\&=\sum_{x_i's=0,\,x_i\neq x_{ji},j=1,\ldots,m}^{1}P(X_1=x_1)\ldots P\left(X_{i_m}=x_{i_m}\right)\\&=P\left(X_{i_1}=x_{i_1}\right)\ldots P\left(X_{i_m}=x_{i_m}\right)\end{aligned} \tag{A3}$$

Equation (A3) proves that events $X_{i_1} = x_{i_1}, \ldots, X_{i_m} = x_{i_m}$ are independent. Therefore, the sequence of Bernoulli random variables $X_1, \ldots, X_n$ is independent.

Appendix B: Pivoting a CDF

We provide here a detailed proof of Theorem 9.2.14 of Casella and Berger (2002, Chapter 9) for computing the confidence intervals.

THEOREM *Let $X \in \chi$ be a discrete statistic with cdf $F_x(x|\theta) = P(X \leq x|\theta)$. Suppose that for each $X \in \chi$, $\theta_L(x)$ and $\theta_U(x)$ are defined as follows:*

1. *If $F_x(x|\theta)$ is a decreasing function of θ for each x, define $\theta_L(x)$ and $\theta_U(x)$ by*

$$P\big(X \leq x \,|\, \theta_U(x)\big) = \frac{\alpha}{2}, \quad P\big(X \geq x \,|\, \theta_L(x)\big) = \frac{\alpha}{2} \tag{B1}$$

2. *If $F_x(x|\theta)$ is an increasing function of θ for each x, define $\theta_L(x)$ and $\theta_U(x)$ by*

$$P\big(X \geq x \,|\, \theta_U(x)\big) = \frac{\alpha}{2}, \quad P\big(X \leq x \,|\, \theta_L(x)\big) = \frac{\alpha}{2} \tag{B2}$$

Then the interval $[\theta_L(x), \theta_U(x)]$ is a $100 \times (1 - \alpha)\%$ confidence interval for θ.

PROOF OF (1) Assume first that $F_x(x|\theta)$ is a decreasing function of θ for each x. Consider the new random variable $T = F_x(x|\theta)$ and assume that the range of X is an ordered set $\{x_0, x_1, \ldots\}$; then the event $\{T > y\} = \{F_x(x|\theta) > y\}$ is equal to the event $\{X \geq x_n\}$, where $x_n = \inf\{x_i : F_x(x_i|\theta) > y\}$. Therefore,

$$\begin{aligned} P(F_X(X|\theta) > y) &= \sum_{i=n}^{\infty} P(X = x_i|\theta) \\ &= P(X \geq x_n|\theta) \end{aligned} \tag{B3}$$

By definition of x_n, $P(X \geq x_n|\theta) = 1 - F_x(x_{n-1}|\theta) \geq 1 - y$. Therefore,

$$P\big(F_x(X|\theta) > y\big) \geq 1 - y \tag{B4}$$

If y is between x_{n-1} and x_n, then the inequality in (B4) is strict. This implies that $P(F_x(X|\theta) \leq y) \leq y$. Likewise, we can prove that $P(\bar{F}_x(X|\theta) \leq y) \leq y$, where $\bar{F}_x(X|\theta) = P(X \geq x|\theta)$. Therefore, the set

$$A(\theta) = \left\{ x : F_x(x|\theta) \leq \frac{\alpha}{2} \quad \text{and} \quad \bar{F}_x(x|\theta) \leq \frac{\alpha}{2} \right\} \tag{B5}$$

is a rejection region with level of significance α. This implies that (see Casella and Berger (2002)) the set

$$C(x) = \{\theta : x \in A(\theta)\} \tag{B6}$$

is a $100 \times (1-\alpha)\%$ confidence set for the parameter θ. We consider now $\theta_L(x)$ and $\theta_U(x)$ that satisfy the equations

$$F_x(x \mid \theta_U(x)) = \frac{\alpha}{2}, \quad \bar{F}_x(x \mid \theta_L(x)) = \frac{\alpha}{2} \tag{B7}$$

Note that $F_x(x \mid \theta)$ is a decreasing function of θ for each x implies that $\bar{F}_x(x \mid \theta)$ is a non-decreasing function of θ. If $\theta > \theta_U(x)$, then $F_x(x \mid \theta) < \alpha/2$; and if $\theta < \theta_L(x)$, then $\bar{F}_x(x \mid \theta) < \alpha/2$ and

$$\left\{\theta : F_X(X \mid \theta) \le \frac{\alpha}{2} \quad \text{and} \quad \bar{F}_X(X \mid \theta) \le \frac{\alpha}{2}\right\}$$
$$= \{\theta : \theta_L(x) \le \theta \le \theta_U(x)\}$$

Therefore, the $100 \times (1-\alpha)\%$ confidence set for the parameter θ is $C(x) = [\theta_L(x), \theta_U(x)]$.

PROOF OF (2) To prove (2) for the increasing case, we consider the function $h(\theta') = F_x(x \mid -\theta')$, where the parameter $-\theta'$ is moving in the domain of θ. The new function is a decreasing function of θ' and we can use the proof given above. □

In order to apply the former result to the binomial case it is enough to observe that the cdf of the binomial distribution is a decreasing function of the parameter. This derives from the fact that X has a monotone likelihood ratio (MLR), ie, for $\theta_1 < \theta_2$, the likelihood ratio

$$\frac{f_x(x \mid \theta_2)}{f_x(x \mid \theta_1)} \tag{B8}$$

is a decreasing function of x and the following result.

MLR implies a decreasing cdf

PROOF We prove this result for the continuous case without loss of generality. Let $\theta_1 < \theta_2$ and $F_x(x \mid \theta) = P(X \le x \mid \theta)$ be the cdf of X,

$$F_x(x \mid \theta_1) - F_x(x \mid \theta_2) = \int_{\Re} 1_{\{y \le x\}} (f_x(y \mid \theta_1) - f_x(y \mid \theta_2))\, dy$$
$$= \int_{\Re} 1_{\{y \le x\}} R(y) f_x(y \mid \theta_1)\, dy \tag{B9}$$

where

$$R(x) = 1 - \frac{f_x(x \mid \theta_2)}{f_x(x \mid \theta_1)} \tag{B10}$$

is a decreasing function of x from the MLR property. We have now two cases:

1. $$R(x) \geq 0, \quad \frac{f_x(x \mid \theta_2)}{f_x(x \mid \theta_1)} \leq 1, \quad \text{for every } y \leq x$$

2. $$R(x), \quad \frac{f_x(x \mid \theta_2)}{f_x(x \mid \theta_1)}, \quad \text{change sign in the interval } (-\infty, x)$$

The first case implies that $F_x(x \mid \theta_1) - F_x(x \mid \theta_2) \geq 0$. For the second case, consider and $A = \{y \in (-\infty, x) : R(y) > 0\}$ and $B = \{y \in (-\infty, x) : R(y) \leq 0\}$; then, from (B8), we obtain

$$F_x(x \mid \theta_1) - F_x(x \mid \theta_2) = \int_A 1_{\{y \leq x\}} R(y) f_x(y \mid \theta_1)\, dy + \int_B 1_{\{y \leq x\}} R(y) f_x(y \mid \theta_1)\, dy \tag{B11}$$

Because $R(y)$ is decreasing, we have that all points in A have to be smaller than the points in B. Let $1_A(x)$ and $1_B(x)$, the smallest and largest values that $1_{\{y \leq x\}}$ can achieve in A and B, respectively. Then, from (B10), we obtain

$$F_x(x \mid \theta_1) - F_x(x \mid \theta_2) \geq 1_A(x) \int_A R(y) f_x(y \mid \theta_1)\, dy + 1_B(x) \int_B R(y) f_x(y \mid \theta_1)\, dy \tag{B12}$$

We have now that

$$\int_A R(y) f_X(y \mid \theta_1) dy = P_{\theta_1}\left(\frac{f_X(y \mid \theta_2)}{f_X(y \mid \theta_1)} < 1\right) - P_{\theta_2}\left(\frac{f_X(y \mid \theta_2)}{f_X(y \mid \theta_1)} < 1\right) \tag{B13}$$

and

$$\int_B R(y) f_X(y \mid \theta_1) dy = P_{\theta_2}\left(\frac{f_X(y \mid \theta_2)}{f_X(y \mid \theta_1)} < 1\right) - P_{\theta_1}\left(\frac{f_X(y \mid \theta_2)}{f_X(y \mid \theta_1)} < 1\right) \tag{B14}$$

Then (B11) becomes

$$F_X(x \mid \theta_1) - F_X(x \mid \theta_2) \geq \left[1_A(x) - 1_B(x)\right] P_{\theta_1}\left(\frac{f_X(y \mid \theta_2)}{f_X(y \mid \theta_1)} < 1\right) - P_{\theta_2}\left(\frac{f_X(y \mid \theta_2)}{f_X(y \mid \theta_1)} < 1\right) \tag{B15}$$

Then,

$$P_{\theta_2}\left(\frac{f_X(y|\theta_2)}{f_X(y|\theta_1)}<1\right) = \int_{\left(\frac{f_X(y|\theta_2)}{f_X(y|\theta_1)}<1\right)} f_X(y|\theta_2)\mathrm{d}y$$

$$= \int_{\left(\frac{f_X(y|\theta_2)}{f_X(y|\theta_1)}<1\right)} \frac{f_X(y|\theta_2)}{f_X(y|\theta_1)} f_X(y|\theta_1)\mathrm{d}y$$

$$< \int_{\left(\frac{f_X(y|\theta_2)}{f_X(y|\theta_1)}<1\right)} f_X(y|\theta_1)\mathrm{d}y$$

$$= P_{\theta_1}\left(\frac{f_X(y|\theta_2)}{f_X(y|\theta_1)}<1\right) \tag{B16}$$

If $y \in A$ with $y > x$, then $1_A(x) = 1_B(x) = 0$. Otherwise, $1_A(x) = 1$ and $1_B(x) = 0$ or $1_A(x) = 1_B(x) = 1$. This implies that $1_A(x) \geq 1_B(x)$. This fact, together with (B15) and (B16), results in the inequality

$$F_x(x|\theta_1) \geq F_x(x|\theta_2) \tag{B17}$$

This implies that the cdf of X is a decreasing function of the parameter θ.

REFERENCES

Basel Committee on Banking Supervision (1996). Supervisory framework for the use of "back testing" in conjunction with the internal models approach to market risk capital requirements. Bank for International Settlements, Basle, January.

Brown, L., Cai, T., and DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science* **16**(2), 101–33.

Casella, G., and Berger, R. (2002). *Statistical inference*, Second edition. Duxbury Advance Series, Pacific Grove, CA 93950, USA.

Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review* **39**, 841–62.

Crnkovic, C., and Drachman, J. (1996). Quality control. *Risk* **9** (September), 139–43.

Jorion, P. (2000). *Value at risk: the new benchmark for managing financial risk*, Second edition. McGraw-Hill, New York.

Fallon, W. (1996). Calculating value-at-risk. Working paper presented at the Wharton Financial Institution Center's Conference on Risk Management in Banking.

Haas, M. (2001). New methods in backtesting. Working paper, Financial Engineering Research Center caesar, Bonn.

Kupiec, P. (1995). Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives*, Winter, 73–8.

Lopez, J. A. (1996). Regulatory evaluation of value-at-risk models. Working paper presented at the Wharton Financial Institution Center's Conference on Risk Management in Banking.